2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Cross Scene Prediction via Modeling Dynamic Correlation using Latent Space Shared Auto-Encoders

Shaochi Hu, Donghao Xu, *Member, IEEE,* and Huijing Zhao, *Member, IEEE*

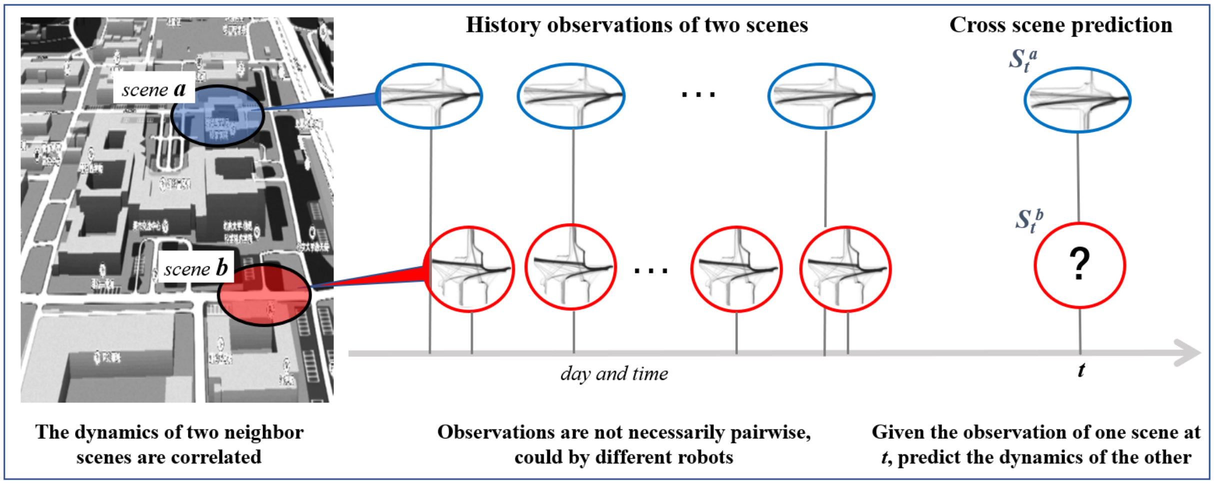

scene *a*
scene *b*
History observations of two scenes
Cross scene prediction
S_t^a
S_t^b
?
day and time
t
The dynamics of two neighbor scenes are correlated
Observations are not necessarily pairwise, could by different robots
Given the observation of one scene at *t*, predict the dynamics of the other

Fig. 1: Can we make cross-scene prediction via modeling the correlations of scene dynamics on unsynchronized history observations?

***Abstract*— This work addresses on the following problem: given a set of unsynchronized history observations of two scenes that are correlative on their dynamic changes, the purpose is to learn a cross-scene predictor, so that with the observation of one scene, a robot can onlinely predict the dynamic state of the other. A method is proposed to solve the problem via modeling dynamic correlation using latent space shared auto-encoders. Assuming that the inherent correlation of scene dynamics can be represented by shared latent space, where a common latent state is reached if the observations of both scenes are at an approximate time. A learning model is developed by connecting two auto-encoders through the latent space, and a prediction model is built by concatenating the encoder of the input scene with the decoder of the target one. Simulation datasets are generated imitating the dynamic flows at two adjacent gates of a campus, where the dynamic changes are triggered by a common working and teaching schedule. Similar scenarios can also be found at successive intersections on a single road, gates of a subway station, etc. Accuracy of cross-scene prediction is examined at various conditions of scene correlation and pairwise observations. Potentials of the proposed method are demonstrated by comparing with conventional end-to-end methods and linear predictions.**

I. INTRODUCTION

The ability to deal with dynamic change of environment is important for robots to achieve lifelong and robust autonomy. Map-based localization approaches could fail if the map is far different from the current environment, and planning can be harder without the knowledge of the dynamic environment.

Some methods are proposed to model the dynamic change of the environment in different aspects. The basic idea is maintaining a database that saves all different observations of the environment, and localization is performed in all past maps [1]. However, this is only a kind of data collection without the modeling process for the change of environment. And with the increasing of the database, computation efficiency and localization in real-time are rapidly influenced. In some scenarios, the change of environment is periodic, which inspires frequency map approach that models the dynamic environment as the sum of some periodic functions [2]. This is a signal level modeling and is able to predict the future of the environment. In general, the dynamic change between neighboring scenes are related, such as traffic flow changing between intersections, and learning the relationships between them is another kind of modeling. Nicholas [3] applies mutual information based method to learn the temporal observability relationships between them.

This work addresses on a new problem: can we make inference by modeling the correlations of scene dynamics on history observations? As illustrated in Fig. 1, two scenes are adjacent, such as nearby gates of the campus, successive intersections on a single road, gates of a subway station, etc. Dynamic changes of the two scenes are correlated, which are triggered by some common events, such as working and teaching schedules of the campus, the period of the

This work is supported by the National Natural Science Foundation of China (61973004) and High-performance Computing Platform of Peking University.

S.Hu, D.Xu and H.Zhao are with the Peking University, with the Key Laboratory of Machine Perception (MOE), and also with the School of Electronics Engineering and Computer Science

Correspondence: H. Zhao, zhaohj@cis.pku.edu.cn.

978-1-7281-6211-9/20/$31.00 ©2020 IEEE

traffic signal, the train's pit stop and so forth. There have been history observations of both scenes, whereas they are not synchronized as they could be measured by different robots, i.e. the observations are not necessarily in pairs. At a certain time, given the observation of one scene, we want to predict the dynamic state of the other. This research proposes a method of cross-scene prediction via modeling dynamic correlation using latent space shared auto-encoders, which is developed based on an assumption that the inherent correlation of scene dynamics can be represented by shared latent space, and a common latent state is reached if the observations of both scenes are at an approximate time. A learning model is thus developed by connecting two auto-encoders through the latent space, and a prediction model is built by concatenating the encoder of the input scene with the decoder of the target one. Simulation datasets are generated, where two scenes are designed imitating the dynamic flows at two adjacent gates of Peking Univ., and a simulator is developed to obtain scene maps for hours. Accuracy of cross-scene prediction is examined, and the performance at various conditions of scene correlation and pairwise observations are elaborated. Potentials of the proposed method are demonstrated by comparing with conventional end-to-end methods and linear predictions.

This paper is organized as follows. The related works are reviewed in SectionII. SectionIII explains the details of our method. SectionIV and SectionV show implementation details of model and simulation. Experimental results are in SectionVI.

II. RELATED WORKS

There are some researches focus on how to model or predict the dynamic change of the environment. Spectrum-analysis based methods [4] [2] discretize environment into binary voxels indicating they are occupied or not, and model each voxel as the sum of a series of periodic signals by the frequency spectra of observed data. Their ability to predict environment improves localization accuracy [5] [6] and efficiency of map updating [7] [8]. Some methods apply long-term and short-term memory in dynamic scene mapping to remove nonexistent features and increase emerging features [9] [10] [11]. The bag-of-word method is also used to predict images between seasons [12]. Mutual information based method [3] predicts images of neighboring scenes by calculating the correlation of collected data, this work is similar to ours but the essence is different because it only learns the temporal relationship between data without considering what makes data correlate or modeling the essence.

In this paper, the dynamic of a scene is caused by moving objects like pedestrians, and there are lots of studies about traffic behavior and scene modeling in the surveillance field. There is a shift from detecting-and-tracking of vehicle state and defining interested events towards machine learning-based approaches to automatically extract meaningful pattern [13]. Similar trajectories are clustered to model structure or path of scene [14] [15]. Topic model based methods convert conceptions of natural language processing into traffic behavior, and LDA [16] [17]/HDA [18] [19] approaches achieve good results in scene modeling without accurate tracking. Scene modeling methods in the surveillance field are mainly used for abnormal events detection or scene semantic understanding [20], but there are few predictions for the future of the full scene. Besides, they do not consider the correlation between neighboring scenes. We can learn some methods in this field, but our conception of scene modeling is essentially different from theirs.

This work makes an attempt to model the dynamic correlation between neighboring scenes on simulation datasets. In order to quantify how the correlation influences our algorithm, we generate datasets with different correlation coefficient between scenes. Training data with different pairwise observations are randomly sampled to simulate robots data acquisition situation in the true world.

III. METHODOLOGY

A. Problem definition

As illustrated in Fig.1, a and b are two neighbor scenes such as adjacent gates of a campus or consecutive intersections on a single road, where the scene dynamics are strongly correlated. Let $\mathbf{S}^a = \{< S_1^a, t_1^a >, ..., < S_n^a, t_n^a >\}$ and $\mathbf{S}^b = \{< S_1^b, t_1^b >, ..., < S_m^b, t_m^b >\}$ be the history observations of the two scenes. S_i^k denotes the ith observation of scene k at time t_i^k, which can be a grid map that represents the dynamic state of the scene. The observations of both scenes are not necessarily pairwise in time, i.e. $\{t_1^a, ..., t_n^a\} \neq \{t_1^b, ..., t_m^b\}$, as they could be obtained independently by different robots.

The purpose of this work is to learn a predictor $\mathbf{F}$ on $\mathbf{S}^a$ and $\mathbf{S}^b$ by addressing the correlation of scene dynamics, where given the observation S^k of one scene at the current time t, predict the dynamic state of the other, e.g.

$$\hat{S}^b, t = \mathbf{F}(S^a, t) \quad (1)$$

$$\hat{S}^a, t = \mathbf{F}(S^b, t) \quad (2)$$

The formulations can be easily extended to define the problems involving three or more scenes.

B. Modeling dynamic correlation using latent space shared auto-encoders

Assumes that there exists a latent space $\mathbf{Z}$ that records the inherent correlation of scene dynamics at a and b, after encoding the observations $\mathbf{S}^a$ and $\mathbf{S}^b$ individually to the latent space $\mathbf{Z}$,

$$Z_i^a = \mathbf{E}_a(S_i^a) \quad (3)$$

$$Z_j^b = \mathbf{E}_b(S_j^b) \quad (4)$$

S_i^a and S_j^b may share a common state, i.e.

$$\Delta Z = ||Z_i^a - Z_j^b||_2^2 \rightarrow 0$$

if they are the observations of an approximate time, i.e.

$$\Delta t = dis(t_i^a, t_j^b) \rightarrow 0$$

where dis is an operator of time difference by addressing the periodic nature of scene dynamics.

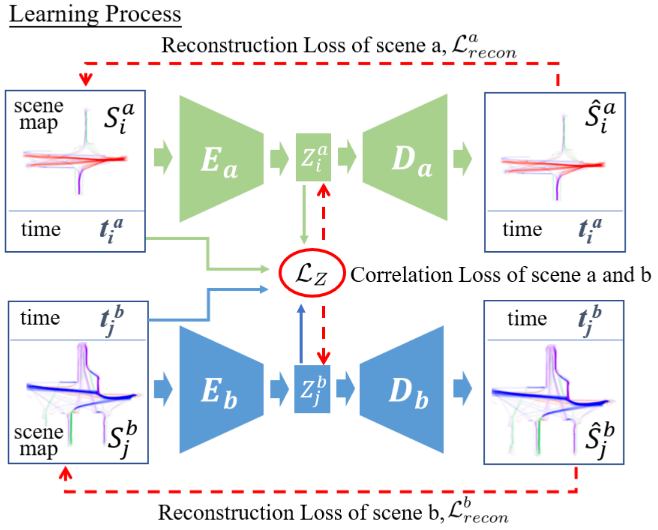


Fig. 2: Modeling dynamic correlation using latent space shared auto-encoders.

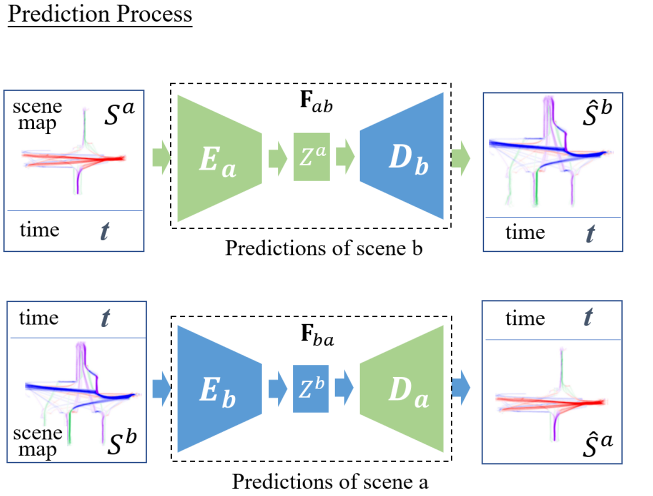


Fig. 3: Cross-scene prediction by concatenating the encoder of the input scene with the decoder of the target one.

As illustrated in Fig.2, the procedure is modeled by combining the auto-encoder structures in this work. Given a pair of history observations of both scenes S_i^a and S_j^b that are measured at t_i^a and t_j^b respectively, each scene map is processed individually through the corresponding encoding-decoding path of the scene.

$$Z_i^a = \mathbf{E}_a(S_i^a), \hat{S}_i^a = \mathbf{D}_a(Z_i^a) \tag{5}$$

$$Z_j^b = \mathbf{E}_b(S_j^b), \hat{S}_j^b = \mathbf{D}_b(Z_j^b) \tag{6}$$

Two reconstruction losses $\mathcal{L}^a_{recon}$ and $\mathcal{L}^b_{recon}$ are defined to evaluate the auto-encoder's accuracy of each scene,

$$\mathcal{L}^a_{recon} = \left\| S_i^a - \hat{S}_i^a \right\|_2^2 \tag{7}$$

$$\mathcal{L}^b_{recon} = \left\| S_j^b - \hat{S}_j^b \right\|_2^2 \tag{8}$$

and a correlation loss is defined to constrain equivalent latent states if the scene dynamics are observed at approximative time points. c is a constant.

$$\mathcal{L}_Z = \exp(-c \cdot \Delta t) \cdot \left\| Z_i^a - Z_j^b \right\|_2^2 \tag{9}$$
$$\Delta t = dis(t_i^a, t_j^b)$$

Therefore, model learning is conducted by optimizing the following total loss

$$\min_{E_a,E_b,D_a,D_b} \mathcal{L}^a_{recon} + \mathcal{L}^b_{recon} + \lambda \mathcal{L}_Z \tag{10}$$

where λ is a hyperparameter that is assigned 0.1 in this research.

The prediction model $\mathbf{F}$ is built by concatenating the encoder of the input scene with the decoder of the target one, as illustrated in Fig.3. For example, at the current time t, given the observation S^a of scene a, the dynamic state of scene b can be predicted by

$$\hat{S}^b, t = \mathbf{F}_{ab}(S^a, t) \tag{11}$$

$$\mathbf{F}_{ab} = E_a o D_b \tag{12}$$

and vice versa

$$\hat{S}^a, t = \mathbf{F}_{ba}(S^b, t) \tag{13}$$

$$\mathbf{F}_{ba} = E_b o D_a \tag{14}$$

IV. IMPLEMENTATION DETAILS

A. Network design

As illustrated in Fig.4, the network contains two autoencoders that have the same structure. We use PyTorch framework to realize the autoencoder [21], which is composed of convolutional, fully connected and upsample layers.

There is no pooling layer in the encoder part, and input size is reduced only by convolutional layers with stride=2. For the decoder part, we use ×2 upsampling with same-padding convolutional layers to extend the size of the input, instead of deconvolutional layers. Such a structure can make the network retain more information.

In encoders, the input size changes from $512 \times 512 \times 4$ to $64 \times 64 \times 8$ by 3 Conv2d layers, and then is reduced to 2 dimensions(the latend variable Z) by 2 FC layers. In decoders, the size is extented from 2 to $64 \times 64 \times 8$ by 3 FC

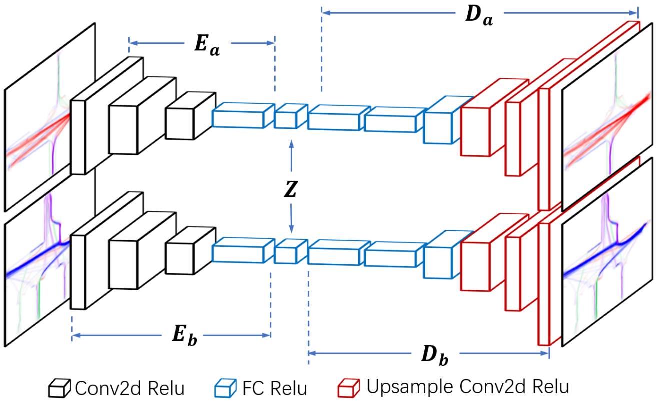
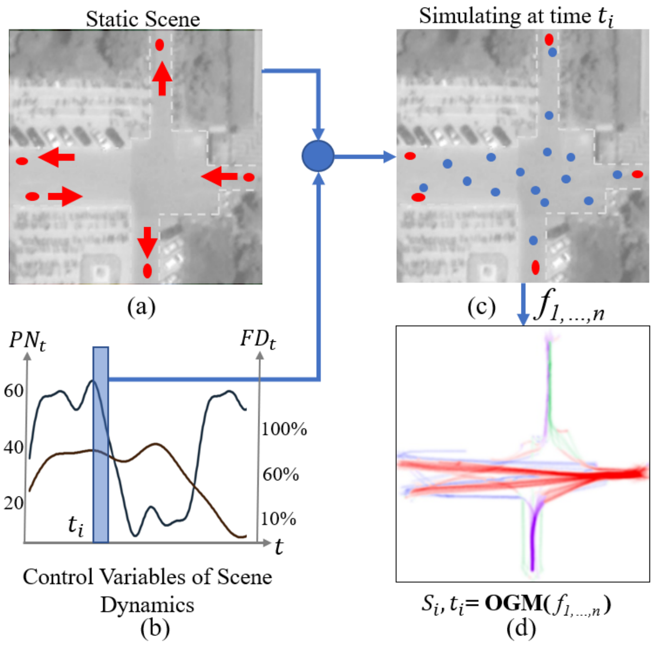


Fig. 4: The network structure of autoencoders.



(a)

(c)

Control Variables of Scene Dynamics
(b)

S_i, t_i= **OGM**($f_{1,\dots,n}$)
(d)

Fig. 5: Simulation pipeline. (a)scene layout, (b)series of control variables of the dynamic flows, (c) a simulation frame of the dynamic objects(blue points), (d) a scene map on a frame sequence during a short time window.

layers, and then restores to $512 \times 512 \times 4$ by 3 Upsample and Conv2d layers.

B. Scene map

A grid map is used to represent the dynamic state of a scene, where each pixel is a four-dimensional vector, recording the number of dynamic objects passing through the location during a short time window τ on four discretized directions. In this research, the map has a dimension of 512×512 and a pixel size of 0.2 meters, $\tau = 1$ min, and the four directions correspond to the East, West, South, and North in the world coordinate system. In this paper, pixels of scene map are visualized by the most dominant flow crossing the pixels at the time, where red, blue, purple and green represent the four discretized directions to the west, east, south and north, respectively, the brighter the color, the higher the dynamic flow.

V. SIMULATION DATASETS

A simulator is developed to generate simulation datasets for experiments. The simulation pipeline is shown in Fig.5. Without loss of generality, we assume that each scene has its inherent structure of the dynamic flows that connect a set of entrance and exit points of the scene, shown as red points in Fig.5(a), whereas the volume of each flow may change with time due to some underlying events. Therefore, time series $\mathcal{C}$ of a set of control variables are designed as illustrated in Fig.5(b) to guide the simulation of dynamic objects. In this research, two control variables are designed, which are the total **p**eople **n**umber PN_t and the main **f**low **d**irection FD_t. Two main flow directions are defined, where FD_t is the percentage of people entering the campus, leaving the rest $1 - FD_t$ going out. At a time t, if the total people number at the frame is less than PN_t, new people are generated to meet the insufficient number. Among the new people, FD_t are generated at the entrance point of the gate following a randomly chosen flow entering the campus, while $1 - FD_t$ are generated randomly at the start point of a flow going out of the campus.

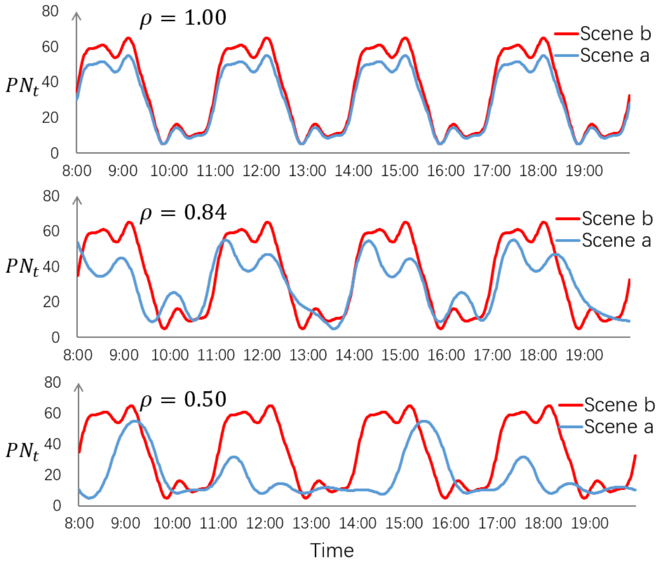


Fig. 6: Scene dynamic correlation is simulated by designing correlative time series of control variables. Three patterns are designed with different correlation coefficient ρ of the time series on the control variable PN_t.

People flows are simulated by referring to Helbing's work [22]. Each scene map is estimated on a sequence of simulation frames as

$$S_i, t_i = \mathbf{OGM}(f_{1,\dots,n}) \tag{15}$$

In this research, simulation frames $f_{1,\dots,n}$ are generated at 10Hz. Each scene map represents the dynamic state during a short time window of $\tau = 1$ min, therefore $n_f = 600$ frames are used to estimate a S_i at t_i.

Two scenes are simulated by imitating the dynamic flows at two adjacent gates of Peking Univ., which are triggered by almost the same events, e.g. working and teaching schedules of the campus. Similar scenarios can also be found at adjacent intersections on a single road, subway stations, gates of a stadium, etc. Therefore, correlated time series of control variables at both scenes are designed as shown in Fig. 6. Three kinds of patterns are designed with the correlation coefficients ρ =1.0, 0.84 and 0.5 of PN_t of two scenes, representing the strong, middle and less correlative scenes. Here people number PN_t of two scenes are designed to control the correlation of two scenes, and we keep the main flow direction FD_t of two scenes the same.

Following each pattern of time series in Fig.6, a simulation is conducted from 8:00 to 20:00, where 720 scene maps are generated every 1 minute for both scenes. Part of scene maps are selected to simulate the different data acquisition situation, which have α =0%,31%,72%,100% of

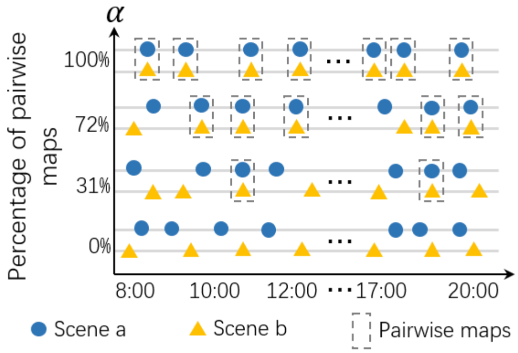


Fig. 7: Datasets generation of various percentage of pairwise scene maps, α =0%,31%,72%,100%.

pairwise observations as shown in Fig.7. Therefore, a total of 12 datasets containing three correlation patterns and four percentages of pairwise observations are generated, which are used in the experiments. In each particular experiment, although the proposed and baseline methods are trained and test on the same dataset, the number of scene maps used in training could be different due to the requirements on pairwise observations of each method, which is detailed in Tab. I.

TABLE I: The number of maps in training and testing of the cross-scene prediction models

Methods / Datasets		$\mathbf{F}_{our}$	$\mathbf{F}_{E2E}$	$\mathbf{F}_{E2E\Delta t}$	$\mathbf{F}_{linear}$
Training	$\alpha = 0$	72	0	72	-
	$\alpha = 31\%$	72	22	72	-
	$\alpha = 72\%$	72	51	72	-
	$\alpha = 100\%$	72	72	72	-
Testing	-	36	36	36	36

VI. Experimental Results

A. Evaluation measures

1) Prediction error: Given two maps S_1 and S_2 of size $W \times H \times C$, mean square error(MSE) is used to measure the difference between them

$$\mathcal{D}_s(S_1, S_2) = \frac{1}{W \times H \times C} \sum_{W,H,C} (S_1 - S_2)^2 \tag{16}$$

Subsequently, for a predicted map $\hat{S}$ with a ground truth S, the prediction error $\mathcal{E}_s$ is defined as

$$\mathcal{E}_s(\hat{S}) = \mathcal{D}_s(\hat{S}, S) \tag{17}$$

2) Dataset variance: A scene map describes the dynamic state of a scene, which is generated by taking statistics on the data frames during a short time window around the time, i.e. $n_f = 600$ frames during $\tau = 1$ min in this research. A scene map has the nature of randomness due to uncontrollable scene dynamics and the method of time windowing, the variance of such randomness is an important reference to prediction accuracy.

Given any series $\mathcal{C}$ of control variables, simulations are conducted for n times. At each sampled time t corresponding to frame number i_t, a time window $[i_0, i_0 + n_f]$ is randomly chosen for m times with $i_0 \in [i_t - n_f, i_t]$, and a scene map is subsequently generated on data frames $f_{i0,\ldots,i_0+n_f}$. Therefore, $n * m$ scene maps $\{S_1, \ldots S_{n*m}\}$ are generated, and inherent variance $\mathcal{V}_s(\mathcal{C}, t)$ of scene map for $\mathcal{C}$ and t is estimated below.

$$\mathcal{V}_s(\mathcal{C}, t) = \frac{1}{n * m} \sum_{i=1}^{n*m} \mathcal{D}_s(S_i, \overline{S}) \tag{18}$$

where, $\overline{S} = \frac{1}{n*m} \sum_{i=1}^{n*m} S_i$ is the mean map.

By repeating the above estimations at all sampled time points $t \in \Omega_t$, variance at the level of datasets $\mathcal{V}_d(\mathcal{C})$ can also be found.

$$\mathcal{V}_d(\mathcal{C}) = \frac{1}{|\Omega_t|} \sum_{t \in \Omega_t} \mathcal{V}_s(\mathcal{C}, t) \tag{19}$$

Dataset variance $\mathcal{V}_d(\mathcal{C})$ is the lower bounder of prediction error for any methods, and the closer the prediction error to the dataset variance, the better the result.

B. Baseline methods

1) $\mathbf{F}_{E2E}$ - Conventional end-to-end prediction: By using only pairwise observations in training datasets, a pair of conventional end-to-end predictors $\mathbf{F}_{E2E}$ can be trained as illustrated in Fig. 8 to predict a $\hat{S}_b$ of scene b on S_a of a, and vice versa.

$$\hat{S}^b, t = \mathbf{F}_{E2E,ab}(S^a, t) \tag{20}$$

$$\hat{S}^a, t = \mathbf{F}_{E2E,ba}(S^b, t) \tag{21}$$

2) $\mathbf{F}_{E2E\Delta t}$ - End-to-end prediction with compensation of time difference: However, the observations are not necessarily pairwise, which could be measured by a multi-robot system. Therefore, the pairwise observations in training datasets are limited when $alpha$ =31%, and none when $alpha$ =0% for $\mathbf{F}_{E2E}$, as shown in TABLE I. A pair of conventional end-to-end predictors with compensation of time difference is

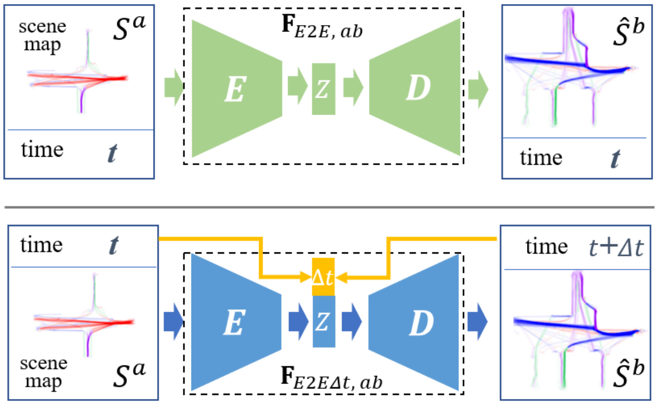


Fig. 8: The baseline methods. Top: conventional end-to-end prediction trained by pairwise maps only. Down: end-to-end prediction with compensation of time difference.

$$\hat{S}^b, t+\Delta t = \mathbf{F}_{E2E\Delta t,ab}(S^a, t, \Delta t) \tag{22}$$

$$\hat{S}^a, t+\Delta t = \mathbf{F}_{E2E\Delta t,ba}(S^b, t, \Delta t) \tag{23}$$

3) $\mathbf{F}_{linear}$ - Linear interpolation: A scene map can also be predicted by finding two history observations of the nearest time by considering the periodic nature of scene dynamics and conducting linear interpolation. Let $S^{\cdot}(t_1)$ and $S^{\cdot}(t_2)$ be the two history observations of the scene at time t_1 and t_2 respectively, and a predicted one of time t is estimated as below.

$$\hat{S}^{\cdot}, t = \frac{S^{\cdot}(t_2) - S^{\cdot}(t_1)}{t_2 - t_1} \times (t - t_2) + S^{\cdot}(t_2) \tag{24}$$

C. Prediction results

We evaluated our method's performance at various conditions of scene correlation(ρ) and pairwise observations(α) comparing with baseline methods, and the quantitative results are shown in TABLE.II. Besides, case study of prediction results is illustrated in Fig.11. Given the input scene maps, the ground truth maps of the other scene are compared with our prediction results, and error maps of ours and baseline methods are also shown on the right four columns. Finally, the study about per map prediction error on the single dataset is exhibited in Fig.12.

1) Prediction accuracy v.s. scene correlation: We explore how the correlation ρ between scenes influences our algorithm, which is taking datasets of the same α but different ρ to experiment. We take datasets with $\alpha = 31\%$ for example.

Quantitative analysis is shown in Fig. 9. When there is high correlation($\rho = 1/0.84$) between scenes, ours(blue) is better than other methods. E2E and E2E$_{\Delta t}$ model have no prior knowledge of scenes but only learn the data mapping of two scenes, and that's why they are worse than ours in high correlation situation. The prediction error of ours increases with the decrease of correlation ρ because the core idea of our method is the latent space of two scenes is shared only when the dynamic changes of scenes are correlated. When the correlation between scenes decreases, the performance is down. And that's why when the scenes are less correlative i.e. $\rho = 0.5$, the prediction error of ours is larger than E2E/ E2E$_{\Delta t}$ methods. The linear prediction model is always the worst. There are the same results for other α shown in TABLE II.

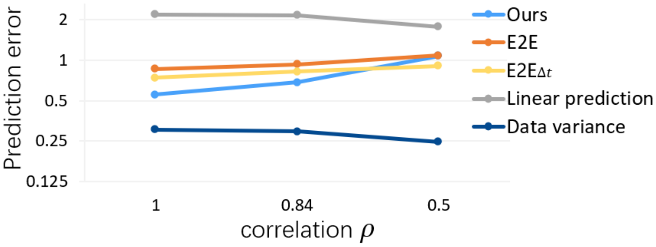


Fig. 9: Average prediction error changes with scene correlation level ρ, a result of α=31%.

Qualitative case study is illustrated in Fig.11(a). Error map A_1 is almost white which means our method achieves good result in high correlation situation. From A_1 to A_3, with the decrease of correlation ρ, the error maps become darker and darker, meaning worse and worse prediction results, and our result A_3 is even worse than $E2E_{\Delta t}$'s result C_3 in less correlation ($\rho = 0.5$) situation.

2) Prediction accuracy v.s. non-pairwise observation: We discuss the influence of percentage α of pairwise data, that is taking datasets of the same ρ but different α to experiment. We take datasets with $\rho = 0.84$ for example.

Quantitative analysis is illustrated in Fig. 10. The prediction error of our method is always the lowest in all percentage α. Ours(blue) and E2E$_{\Delta t}$(yellow) method are not sensitive to if scene maps are pairwise, because the time difference between scene maps is considered in them. The E2E method only processes pairwise data in the training step, so the decrease of pairwise data leads to the reduction of training data, causing the prediction error to raise. And that's also the reason for the lack of results on 0% paired data of E2E method. There are the same results for other correlation ρ in TABLE II.

Qualitative case study is shown in Fig. 11(b). Percentage α does not influence a lot on our method, and the slight difference between prediction error leads to the similar error maps of all methods.

3) Study on single dataset: There are similar results in the study on per map prediction error on single dataset shown in Fig. 12. Our prediction error is close to data variance and always lower than baseline methods through the day when the correlation is strong ($\rho = 1$), and the percentage α of pairwise scene maps rarely influences the performance of our method, shown in Fig. 12(a)&Fig. 12(b). But when there is less correlation ($\rho = 0.5$) between scenes, ours sometimes can be worse than baseline methods, shown in Fig. 12(c)&Fig. 12(d). Finally, Fig. 12(e) & Fig. 12(f) are the people number of one day, and the data variance changes with it. This is because in our pedestrian simulator, every pedestrian's movement is influenced by its nearby people, and when there are lots of people in the scene, the randomness of pedestrians' movement increases, leading to the raising of data variance.

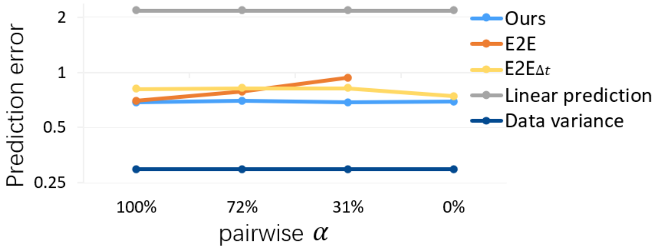


Fig. 10: Average prediction error changes with the percentage of pairwise observations α, a result of ρ=0.84.

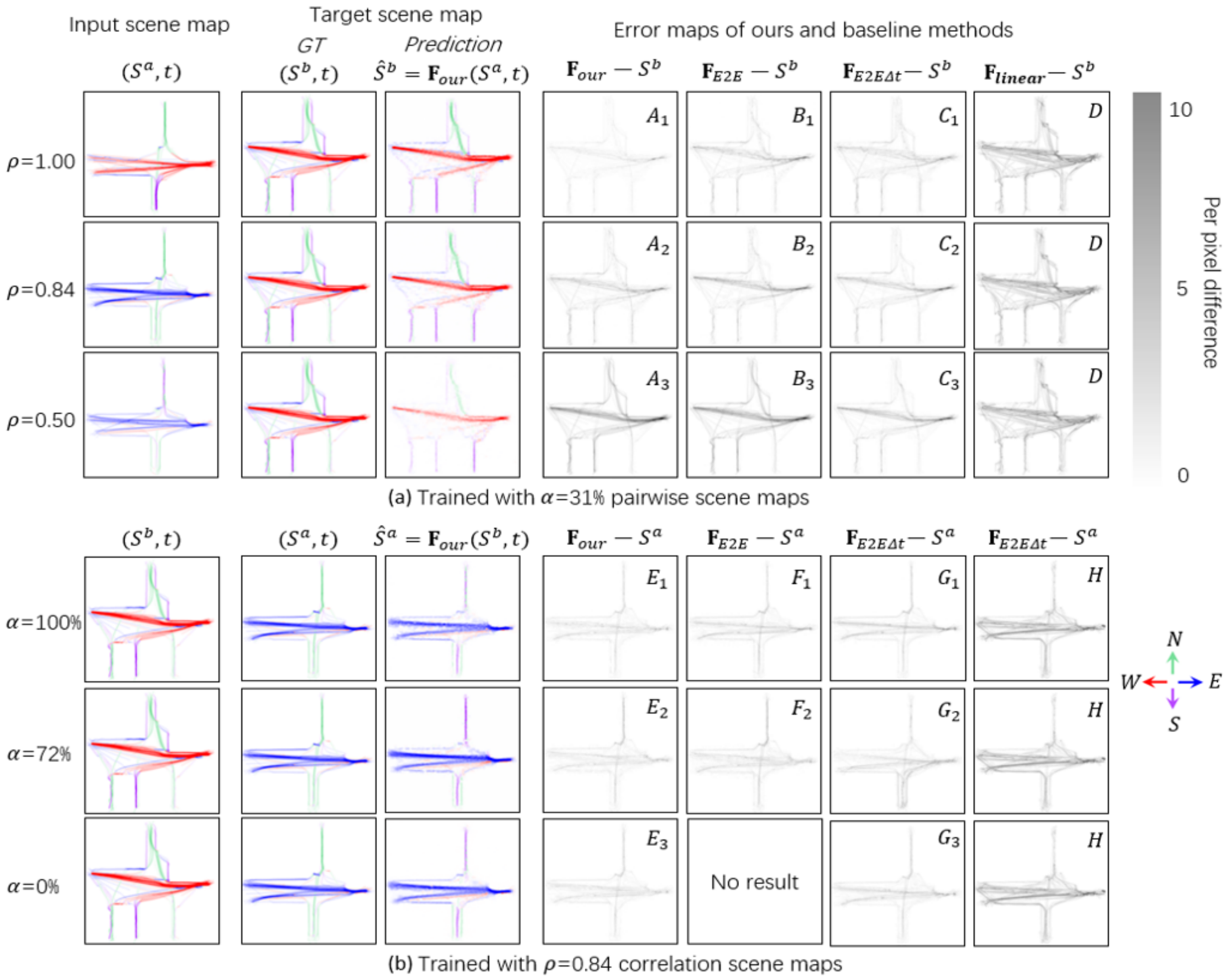


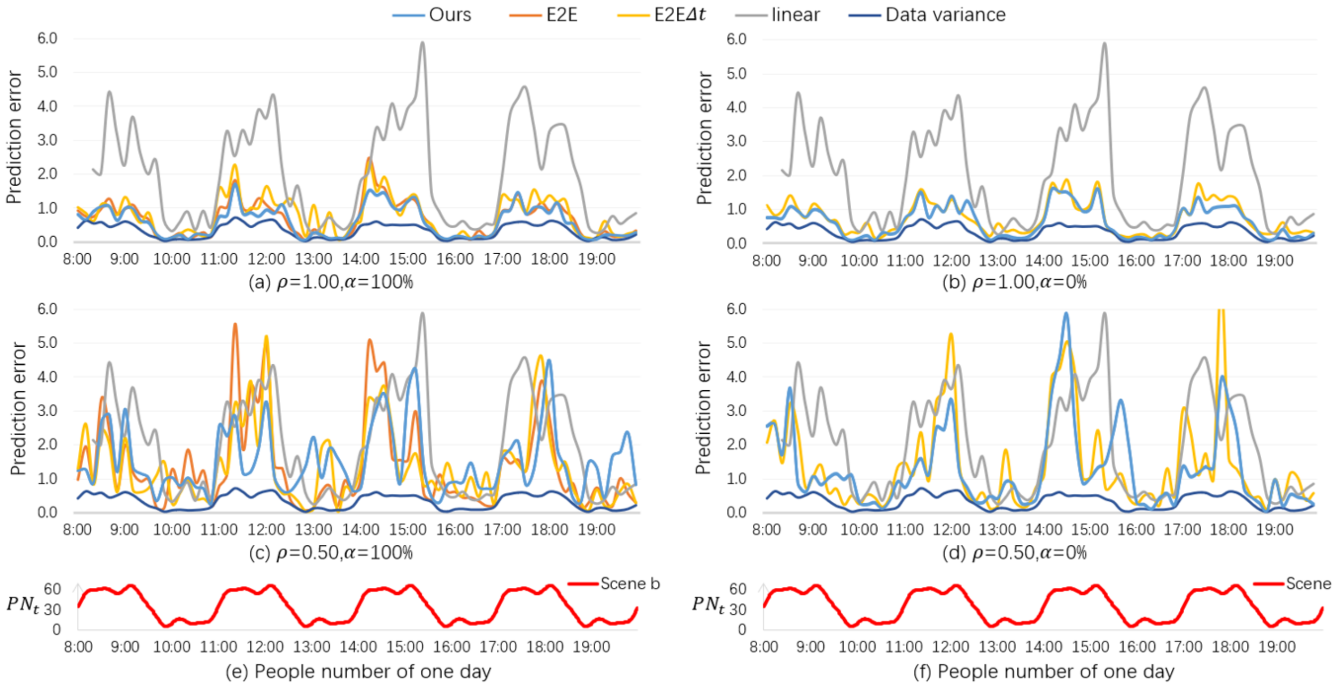
Fig. 11: Case study of prediction results, comparison with baseline methods at various conditions of scene correlation (ρ) and pairwise observations (α).

TABLE II: Average prediction error on each dataset corresponding to a pair of ρ and α

Datasets / Methods	ρ	1.00				0.84				0.50			
	α	100%	72%	31%	0%	100%	72%	31%	0%	100%	72%	31%	0%
Ours		0.549	0.581	0.558	0.569	0.685	0.698	0.686	0.693	1.037	0.974	1.080	0.925
E2E		0.602	0.660	0.864	-	0.702	0.784	0.933	-	0.894	0.938	1.087	-
E2E$_{\Delta t}$		0.716	0.731	0.745	0.700	0.810	0.819	0.820	0.743	0.897	0.876	0.905	0.897
Linear prediction		2.180				2.162				1.777			
Dataset variance		0.305				0.296				0.248			

VII. CONCLUSIONS

This paper is the first try to answer the question: can we make inference by modeling the correlations of scene dynamics on history observations? We formulate the problem as given a set of unsynchronized history observations of two scenes that are correlative on their dynamic changes, learn a cross-scene predictor, where with the observation of one scene, a robot can onlinely predict the dynamic state of the other. The problem is solved by developing a method by modeling the inherent correlation of scene dynamics using latent space shared auto-encoders, where a learning model is established by connecting two auto-encoders through the latent space, and a prediction model is built by concatenating the encoder of the input scene with the decoder of the target one. The method is examined through simulation, where the dynamic flows at two adjacent gates of campus are imitated. The problem is adaptive to other scenarios such as successive intersections on a single road, gates of subway stations, etc., where the dynamic changes are triggered by some common events. Cross-scene prediction accuracy is



(a) ρ=1.00,α=100%

(b) ρ=1.00,α=0%

(c) ρ=0.50,α=100%

(d) ρ=0.50,α=0%

(e) People number of one day

(f) People number of one day

Fig. 12: Per map prediction error on each dataset corresponding to a pair of ρ and α.

examined at various conditions of scene correlation and pairwise observations, and the results show that the proposed method can better solve the problem than the conventional end-to-end and linear predictions ones. Future work will be addressed on real-data collection and processing, and the inference on dynamic correlations of more adjacent scenes will also be studied.

References

[1] W. Churchill and P. Newman, “Experience-based navigation for long-term localisation,” *International Journal of Robotics Research*, vol. 32, no. 14, pp. 1645–1661, 2013.

[2] T. Krajnik, J. P. Fentanes, J. M. Santos, and T. Duckett, “FreMEn: Frequency map enhancement for long-term mobile robot autonomy in changing environments,” *IEEE Transactions on Robotics*, vol. 33, no. 4, pp. 964–977, 2017.

[3] N. Carlevaris-Bianco and R. M. Eustice, “Learning temporal co-observability relationships for lifelong robotic mapping,” *IROS Workshop on Lifelong Learning for Mobile Robotics Applications*, 2012.

[4] T. Krajník, J. P. Fentanes, G. Cielniak, C. Dondrup, and T. Duckett, “Spectral analysis for long-term robotic mapping,” *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 3706–3711, 2014.

[5] T. Krajnik, J. P. Fentanes, O. M. Mozos, T. Duckett, J. Ekekrantz, and M. Hanheide, “Long-term topological localisation for service robots in dynamic environments using spectral maps,” *IEEE International Conference on Intelligent Robots and Systems*, pp. 4537–4542, 2014.

[6] J. P. Fentanes, B. Lacerda, T. Krajnik, N. Hawes, and M. Hanheide, “Now or later? Predicting and maximising success of navigation actions from long-term experience,” *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2015-June, no. June, pp. 1112–1117, 2015.

[7] J. M. H. Santos, T. Krajnik, J. P. Fentanes, and T. Duckett, “Lifelong Information-Driven Exploration to Complete and Refine 4-D Spatio-Temporal Maps,” *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 684–691, 2016.

[8] T. Krajník, J. M. Santos, and T. Duckett, “Life-long spatio-temporal exploration of dynamic environments,” *2015 European Conference on Mobile Robots, ECMR 2015 - Proceedings*, 2015.

[9] F. Dayoub and T. Duckett, “An adaptive appearance-based map for long-term topological localization of mobile robots,” *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, pp. 3364–3369, 2008.

[10] T. Morris, F. Dayoub, P. Corke, G. Wyeth, and B. Upcroft, “Multiple map hypotheses for planning and navigating in non-stationary environments,” *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 2765–2770, 2014.

[11] J. S. Berrio, J. Ward, S. Worrall, and E. Nebot, “Updating the visibility of a feature-based map for long-term maintenance,” *IEEE Intelligent Vehicles Symposium, Proceedings*, vol. 2019-June, no. Iv, pp. 1173–1179, 2019.

[12] P. Neubert, N. Sunderhauf, and P. Protzel, “Appearance change prediction for long-term navigation across seasons,” *2013 European Conference on Mobile Robots, ECMR 2013 - Conference Proceedings*, pp. 198–203, 2013.

[13] B. T. Morris and M. M. Trivedi, “Understanding vehicular traffic behavior from video: a survey of unsupervised approaches,” *Journal of Electronic Imaging*, 2013.

[14] D. Makris and T. Ellis, “Learning semantic scene models from observing activity in visual surveillance,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2005.

[15] C. Piciarelli and G. L. Foresti, “On-line trajectory clustering for anomalous events detection,” *Pattern Recognition Letters*, 2006.

[16] S. Kwak, “Detection of dominant flow and abnormal events in surveillance video,” *Optical Engineering*, 2011.

[17] L. Song, F. Jiang, Z. Shi, and A. K. Katsaggelos, “Understanding dynamic scenes by hierarchical motion pattern mining,” in *Proceedings - IEEE International Conference on Multimedia and Expo*, 2011.

[18] X. Wang, X. Ma, and W. E. L. Grimson, “Unsupervised Activity Perception in Crowded and Complicated Scenes Using Hierarchical Bayesian Models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[19] T. S. Haines and T. Xiang, “Delta-dual hierarchical Dirichlet processes: A pragmatic abnormal behaviour detector,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2011.

[20] I. Saleemi, K. Shafique, and M. Shah, “Probabilistic modeling of scene dynamics for applications in visual surveillance,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[21] P. Chakravarty, P. Narayanan, and T. Roussel, “GEN-SLAM: Generative modeling for monocular simultaneous localization and mapping,” in *Proceedings - IEEE International Conference on Robotics and Automation*, 2019.

[22] D. Helbing and P. Molnár, “Social force model for pedestrian dynamics,” *Physical Review E*, 1995.